W522–W525 *Nucleic Acids Research, 2004, Vol. 32, Web Server issue*
DOI: 10.1093/nar/gkh414

PROSPECT-PSPP: an automatic computational pipeline for protein structure prediction

Jun-tao Guo[1], Kyle Ellrott[1], Won Jae Chung[1], Dong Xu[2], Serguei Passovets[3] and Ying Xu[1,3,*]

[1]Department of Biochemistry and Molecular Biology, University of Georgia, Athens, GA 30606, USA, [2]Department of Computer Science, University of Missouri – Columbia, Columbia, MO 65211, USA and [3]Computational Biology Institute, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA

Received February 15, 2004; Revised and Accepted April 2, 2004

ABSTRACT

Knowledge of the detailed structure of a protein is crucial to our understanding of the biological functions of that protein. The gap between the number of solved protein structures and the number of protein sequences continues to widen rapidly in the post-genomics era due to long and expensive processes for solving structures experimentally. Computational prediction of structures from amino acid sequence has come to play a key role in narrowing the gap and has been successful in providing useful information for the biological research community. We have developed a prediction pipeline, PROSPECT-PSPP, an integration of multiple computational tools, for fully automated protein structure prediction. The pipeline consists of tools for (i) preprocessing of protein sequences, which includes signal peptide prediction, protein type prediction (membrane or soluble) and protein domain partition, (ii) secondary structure prediction, (iii) fold recognition and (iv) atomic structural model generation. The centerpiece of the pipeline is our threading-based program PROSPECT. The pipeline is implemented using SOAP (Simple Object Access Protocol), which makes it easier to share our tools and resources. The pipeline has an easy-to-use user interface and is implemented on a 64-node dual processor Linux cluster. It can be used for genome-scale protein structure prediction. The pipeline is accessible at http://csbl.bmb.uga.edu/protein_pipeline.

INTRODUCTION

In the post-genomics era, the major challenge facing the bioinformatics research community is to determine the biological functions of the genes identified through the large-scale sequencing efforts. Although experimental methods can provide high-resolution structural information, they cannot keep up with the production rate of protein sequences due to the limitations of the current technology. Computational structure prediction methods, on the other hand, have made rapid progress in the past several years, as demonstrated in the community-wide experiments on the Critical Assessment of Techniques for Protein Structure Prediction (CASP) (1) and the Critical Assessment of Fully Automated Structure Prediction (CAFASP) (2). These techniques have matured to such a level that they can provide valuable information and insights for proteins whose structures are not known. Protein structure prediction and modeling tools are becoming an essential part of biological research. Like BLAST (3) searching, structure prediction is now among the first tools applied in a biological investigation, for information collection prior to experiments. Computational prediction tools can often help quickly generate hypotheses to guide the design of experiments. The effectiveness of the methods has been demonstrated by a large number of real-life applications (4). Although computational techniques are yet to consistently produce structures of the same quality as those produced by experimental methods, fold recognition techniques (including both threading and homology-based modeling) can often generate backbone structures with an accuracy of 4 Å root mean square distance (RMSD) for a fairly large class of proteins, as demonstrated in the CASP predictions (1). With this level of prediction accuracy, highly useful functional information can be derived. For example, one can predict the residues involved in protein interaction based on the predicted three-dimensional (3D) structure and a comparison between the predicted structures and some known structures, and then mutation experiments on these residues can be designed. In addition, a predicted fold can also suggest possible functions based on the few function categories associated with the fold.

Compared with BLAST searching and interpretation of its results, protein structure prediction is a more complex process, which typically involves application of multiple tools. A large

*To whom correspondence should be addressed. Tel: +1 706 542 9779; Fax: +1 706 542 9751; Email: xyn@bmb.uga.edu number of computational tools for prediction of different aspects of protein structures are available to the biological research community on the Internet. By effectively utilizing these tools and integrating the prediction results, a human expert predictor could make much more accurate predictions than any individual structure prediction tool. However, to achieve this often requires a good understanding of each individual prediction tool by a human predictor. Failing to achieve this, the full capabilities of structure prediction tools are often under-utilized by the biological community. To help remedy this, we have developed a computational pipeline which employs multiple tools and effective strategies to integrate them. To help explain our design rationale, we summarize a typical and effective process of computer-assisted manual prediction of protein structures, which is often used by expert predictors. Before structure prediction through homology modeling or fold recognition, several preprocessing steps may need to be taken. For example, some protein sequences have signal peptides, which should be removed first since they are not involved in the folding process and are not part of the mature protein structure. Membrane proteins have different physicochemical properties from soluble proteins. Therefore different computational techniques are needed for their prediction. Large proteins with multiple structural domains may need to have their sequences partitioned into individual domains before structural prediction. As a key step in protein structure prediction, protein fold recognition can be achieved through sequence comparison methods, such as PSI-BLAST (3), or through structure-based threading methods. The threading methods can reveal more distant relationships than the sequence-based methods, and can provide more information since these methods use both sequence and structural information. After a structural fold is predicted, a 3D structural model can be generated using homology-modeling tools such as MODELLER (5). Then tools such as PROCHECK (6) and WHAT IF (7) can be used to assess the quality of the predicted model. Additional tools can be employed to search for functional hints about a query protein, and to compare and assess the predicted structure model against the collected information.

We have developed PROSPECT-PSPP for protein structure predictions (8) with the intention of integrating all these prediction and analysis steps into one prediction pipeline, in order to make structure prediction tools more accessible to a larger biological research community. We have observed that by integrating the tools, we can improve the prediction performance of individual tools. In addition to these prediction and analysis capabilities, we have also incorporated a great deal of the expert knowledge of human predictors gained through our prediction experience in CASPs and in real-life applications (4,9). This has been done through implementing a set of decision-making and inference rules as part of the pipeline. These rules guide the use of different prediction tools for different scenarios and help interpret the predicted results.

At the implementation level, the pipeline uses SOAP (Simple Object Access Protocol) technology (http://www.w3.org/TR/SOAP/ and http://www.soaplite.com/) for remote procedure calls. SOAP facilitates easy access to and application of individual tools in the pipeline for advanced users, bypassing the central control mechanism of the pipeline, to assemble their own 'pipelines'. Currently the pipeline is implemented on a large cluster with 128 CPUs, making it possible to perform multiple genome-scale applications simultaneously. PROSPECT-PSPP can be accessed through an interactive web interface at http://csbl.bmb.uga.edu/protein_pipeline.

SYSTEMS AND METHODS

Architecture of the pipeline

PROSPECT-PSPP consists of a list of prediction and analysis tools. A pipeline manager has been developed and implemented to control the flow of the prediction process by calling various tools based on the results from previous prediction steps. The pipeline manager, written in Perl, is invoked by a user's request through a web interface. It stores all user inputs, user-chosen prediction parameters and predicted results, intermediate or final, in a MySQL database. The manager dynamically directs the prediction process by calling individual prediction programs based on the information available in the MySQL database. The final prediction results are displayed through a web interface. Figure 1 shows the architecture of the whole pipeline.

The pipeline manager starts a typical prediction process as follows. It first preprocesses the query sequence by calling tools to identify and remove any signal peptides. It will then predict in a triage step whether the query protein is a soluble protein, a single transmembrane-domain protein or a mixture of transmembrane domain/segment and soluble domain. If the query protein does not involve a transmembrane domain or segment, it will proceed to the fold recognition step. Otherwise, it predicts only the secondary structure of the membrane protein. If a close homolog is found in the PDB (Protein Data Bank) (10) using a sequence-based approach, the alignment between the protein and its structural homolog will be used for homology-based modeling to predict an atomic level structure. If the protein has no structural homologs detectable using sequence-based methods, it then performs fold recognition using a threading program. The sequence–structure alignment from threading analysis will be used for homology modeling. We are currently finishing up implementing an expert system for evaluating and refining fold recognition and structure

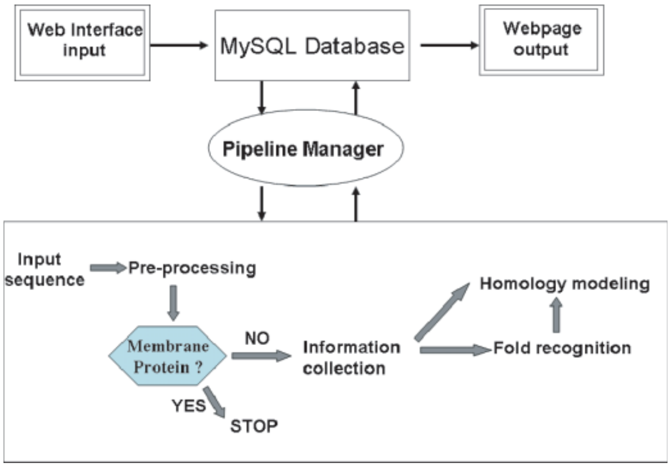

Figure 1. Schematic representation of the protein structure prediction pipeline.

prediction, employing a set of rules (results to be published). This expert system will be added to the manager of the pipeline.

Individual tools and features

The pipeline currently employs the following analysis and prediction tools.

(i) *Pipeline manager*. Currently, the function of the pipeline manger is to make decisions and to control the flow of the prediction process by calling various tools based on the results from previous prediction steps. The manager is implemented as a set of rules.

(ii) *SignalP* (11). SignalP predicts the location of a signal peptide cleavage site, if any, in a protein sequence. PROSPECT-PSPP cuts off any identified signal peptide from the target sequence.

(iii) *ProDom* (12). ProDom is a comprehensive set of protein domain families automatically generated from protein sequence databases. It can be used to identify structural domains in a protein sequence. Since ProDom might return hits with overlapping regions and significantly short regions, our pipeline manager uses a set of rules to make the final prediction of domains.

(iv) *SOSUI* (13). SOSUI is a program for classification of protein types, soluble or membrane; it also makes a prediction of transmembrane helices.

(v) *PSI-BLAST* (3). PSI-BLAST is a popular program for searching sequence similarity. PSI-BLAST in our prediction pipeline is used to search the PDB database for structural homologs. If no structural homologs are identified, an iterative PSI-BLAST will be performed to produce sequence profiles that can be used for fold recognition. PSI-BLAST is also used as part of our secondary structure prediction program.

(vi) *Prospect-SSP*. Prospect-SSP is an in-house program for protein secondary structure prediction. It uses sequence profiles of protein families and a trained neural network for protein secondary structure prediction. The performance of this program (unpublished results) is comparable to other popular secondary structure programs, such as PSI-PRED (14).

Table 1. The confidence index and fold recognition based on combined z-score

z-score intervals	Confidence index	Confidence level	Similarity level
<6	<0.3	unlikely	fold/unrelated
6–8	0.35	low	superfamily/fold
8–10	0.63	medium	superfamily/fold
10–12	0.85	high	family/superfamily
12–20	0.96	very high	family/superfamily
>20	>0.99	certain	family

(vii) *PROSPECT* (15,16). The key component of the pipeline is the fold recognition program, PROSPECT, developed in our lab. PROSPECT is unique in several ways compared with similar programs. It treats pairwise residue contact rigorously using a divide-and-conquer algorithm (15). One of the most important requirements for a protein structure prediction method is its ability to measure the reliability for each prediction, especially for genome-scale predictions. We have developed a confidence index using a combined z-score scheme which measures the reliability of a prediction and a possible structure–function relationship, as shown in Table 1 (8,16). The table shows that a certain range of z-scores corresponds to a different confidence level and structural similarity level and probability of the template being the correct fold. The higher the z-score, the more reliable a prediction is. The structural similarity is based on the SCOP (structural classification of proteins) protein family classification (17). Proteins in the same family have clear evolutionary relationship with high sequence identity (>25%). Two proteins are considered to be in the same superfamily if they have a common evolutionary origin based on their structural and functional features, but with lower sequence identity. If two proteins belong to the same fold category, they do not necessarily have common evolutionary origin. But they do have the same major secondary structure arrangements with the same topological connections.

(viii) *MODELLER* (5). MODELLER is the state-of-art homology-modeling program. It takes sequence–structure alignments as inputs for atomic structure generation. In our pipeline, the alignment files can be generated either by PSI-BLAST searching or by threading using PROSPECT.

A number of other tools are being added to this prediction pipeline, including PROCHECK, WHAT IF for structural quality assessment and an expert system for assessing fold recognition reliability and refining sequence–structure alignments.

Web interface

The pipeline has a web interface providing an interactive environment for users to directly choose prediction tools and specify the parameter values for the tools chosen. Users can also use the default values designed for each tool of the pipeline. After a prediction job is submitted through the web interface, a user can monitor the progress of the prediction process, the status of individual tools and the prediction results. If a prediction process is stopped for whatever reason (e.g. computer failure), the user can resume the prediction process from where it stopped through the web interface. The web interface is implemented using PHP (http://www.php.net). All results are dynamically generated for each user. Therefore, a user can have all submitted jobs displayed on one page.

Systems and implementation

Each prediction tool in the prediction pipeline is implemented as an individual module written in Perl, which provides an interface to the pipeline manager. We used SOAP, a new open web communication standard through XML, for dispatching work to remote machines. For tools implemented with SOAP, the SOAP client in each module calls the SOAP server. The server then dispatches the incoming requests to the machine where the corresponding computational job is being carried out. SOAP is built on the HTTP and XML standards, which makes it an excellent technology for sharing and accessing computational resources using various languages. Advanced users can write SOAP client codes using Java, Perl or other programming languages to access the resources implemented with SOAP. We believe this technology will serve to provide better solutions for interoperability of bioinformatics databases and tools.

A local MySQL database is used to store and organize all the input parameters, prediction status and prediction results. All relevant data is stored in the database in a journal-like fashion, so that if a process dies in the middle of a calculation, its calculation can be resumed from where it stopped. The pipeline manager communicates with MySQL periodically to check the status of each process.

Genome-scale predictions using pipeline

We have carried out a structure prediction for all the open reading frames (ORFs) of *Pyrococcus furiosus*, a target organism in the SouthEast Collaboratory for Structural Genomics (SECSG, http://www.secsg.org/). *P.furiosus* is found in the marine sand surrounding sulfurous volcanoes and can grow at temperatures above 100°C. It can utilize peptides, proteins and some carbohydrates, for example, starch and maltose, as carbon sources. The entire genome of *P.furiosus* is about 2 Mb in length and has 2195 annotated ORFs.

Using the default parameters, we sent all the 2195 ORFs of *P.furiosus* to PROSPECT-PSPP. For the current application, we are mainly interested in protein fold prediction. Hence, we used an option of the pipeline not to generate atomic level structures. Using this option, it takes about 15 min on average to process one protein on a single Linux processor (XEON 3.06 GHz). The prediction for all 2195 ORFs was done in about 12 h on our Linux cluster. Table 2 provides a summary of the prediction results. Out of a total of 2195 ORFs, 540 are predicted as membrane proteins, and 573 have close structural homologs in the PDB detected by PSI-BLAST using a default E-value of 10^{-4}. A total of 190 ORFs have structure predictions with a PROPECT z-score >12, indicating that the prediction reliability of these predictions is at least 96% (16). An additional 84 ORFs have PROSPECT z-score between 6 and 12. The prediction results can be accessed through the Internet at https://csbl.bmb.uga.edu/protein_pipeline/pyrococcusf.php.

Conclusion

In summary, we have developed a computational pipeline for protein structure prediction called PROSPECT-PSPP. Through its web user interface, PROSPECT-PSPP allows a user to customize which tools to use for particular application, e.g. generating detailed structure coordinates or simply predicting structural homologs. For advanced users, PROSPECT-PSPP provides the capacity for them to access individual tools directly so they can dynamically assemble their own pipelines. The pipeline is running on a Linux cluster with 128 CPUs so it can make structure predictions at genome scale. For each predicted structure, PROSPECT-PSPP gives a value indicating the prediction reliability.

Table 2. A summary of predicted structural folds in *Pyrococcus furiosus*

Total number of ORFs	2195
Membrane proteins	540 (24.6%)
PSI-BLAST hits	573 (26.1%)
PROSPECT ($z \geqslant 12$)	190 (8.7%)
PROSPECT ($6 \leqslant z < 12$)	84 (3.8%)
Total number of structural homologs	847 (38.6%)

ACKNOWLEDGEMENTS

The authors would like to thank Ms Shiming Dong and Mr Abhishek Chugh for their technical help. This work is in part supported by the Office of Biological and Environmental Research, US Department of Energy, under Contract DE/FG-2-04ER63714. It is also funded in part by the US Department of Energy's Genomes to Life program (www.doegenomestolife.org) under the project 'Carbon Sequestration in *Synecococcus sp.*: from Molecular Machines to Hierarchical Modeling' (www.genomes-to-life.org).

REFERENCES

1. Moult,J., Fidelis,K., Zemla,A. and Hubbard,T. (2003) Critical assessment of methods of protein structure prediction (CASP)-round V. *Proteins*, **53**(Suppl. 6), 334–339.
2. Fischer,D., Rychlewski,L., Dunbrack,R.L.,Jr, Ortiz,A.R. and Elofsson,A. (2003) CAFASP3: the third critical assessment of fully automated structure prediction methods. *Proteins*, **53**(Suppl. 6), 503–516.
3. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
4. Xu,D., Baburaj,K., Peterson,C.B. and Xu,Y. (2001) Model for the three-dimensional structure of vitronectin: predictions for the multi-domain protein from threading and docking. *Proteins*, **44**, 312–320.
5. Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.
6. Laskowski,R.A., MacArthur,M.W., Moss,D.S. and Thornton,J.M. (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.*, **26**, 283–291.
7. Vriend,G. (1990) WHAT IF: a molecular modeling and drug design program. *J. Mol. Graph.*, **8**, 52–56, 29.
8. Shah,M., Passovets,S., Kim,D., Ellrott,K., Wang,L., Vokler,I., LoCascio,P., Xu,D. and Xu,Y. (2003) A computational pipeline for protein structure prediction and analysis at genome scale. *Bioinformatics*, **19**, 1985–1996.
9. Xu,D., Crawford,O.H., LoCascio,P.F. and Xu,Y. (2001) Application of PROSPECT in CASP4: characterizing protein structures with new folds. *Proteins*, (Suppl. 5), 140–148.
10. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
11. Nielsen,H., Engelbrecht,J., Brunak,S. and von Heijne,G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.*, **10**, 1–6.
12. Servant,F., Bru,C., Carrere,S., Courcelle,E., Gouzy,J., Peyruc,D. and Kahn,D. (2002) ProDom: automated clustering of homologous domains. *Brief Bioinform.*, **3**, 246–251.
13. Hirokawa,T., Boon-Chieng,S. and Mitaku,S. (1998) SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics*, **14**, 378–379.
14. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.
15. Xu,Y. and Xu,D. (2000) Protein threading using PROSPECT: design and evaluation. *Proteins*, **40**, 343–354.
16. Kim,D., Xu,D., Guo,J.T., Ellrott,K. and Xu,Y. (2003) PROSPECT II: protein structure prediction program for genome-scale applications. *Protein Eng.*, **16**, 641–650.
17. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.